Metagenomics and disease

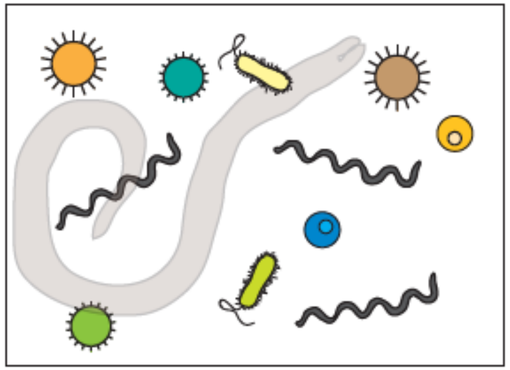
Mark Stenglein, GDW

Genomics of Disease in Wildlife

Metagenomics is the study of >1 genome

Many genomes

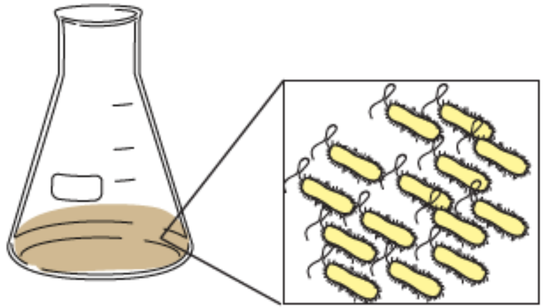
soil community

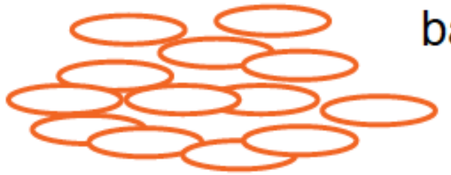


soil ‘metagenome’

1 genome

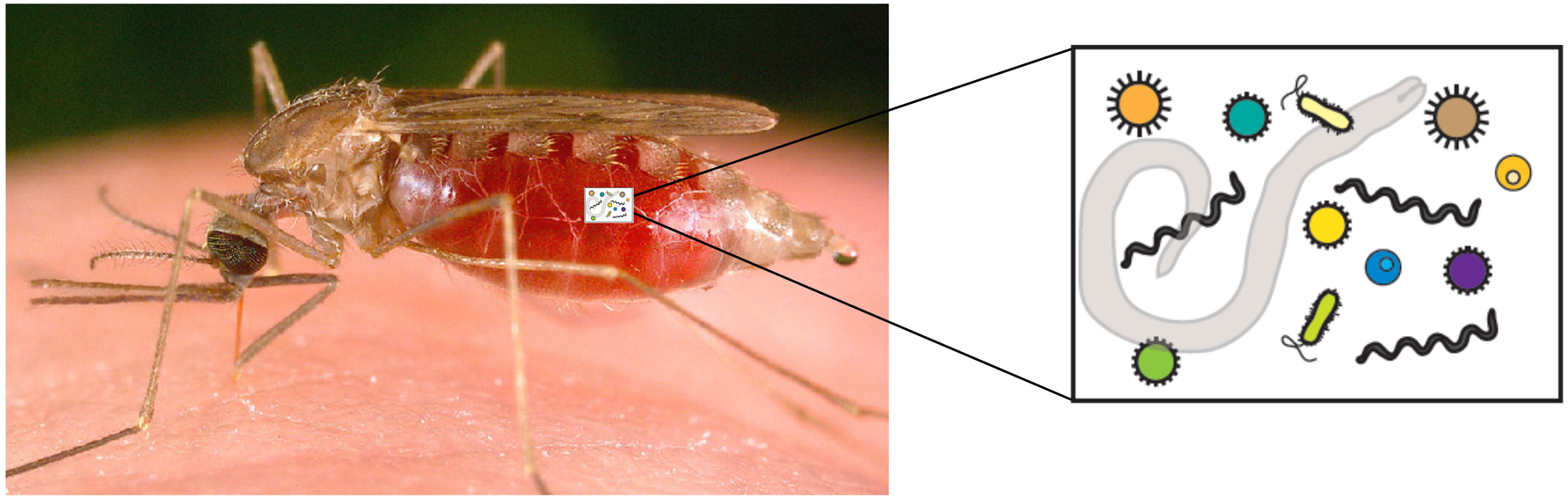
bacterial isolate





bacterial genome

If you sequence total nucleic acid from an intact multicellular organism, you are doing metagenomics

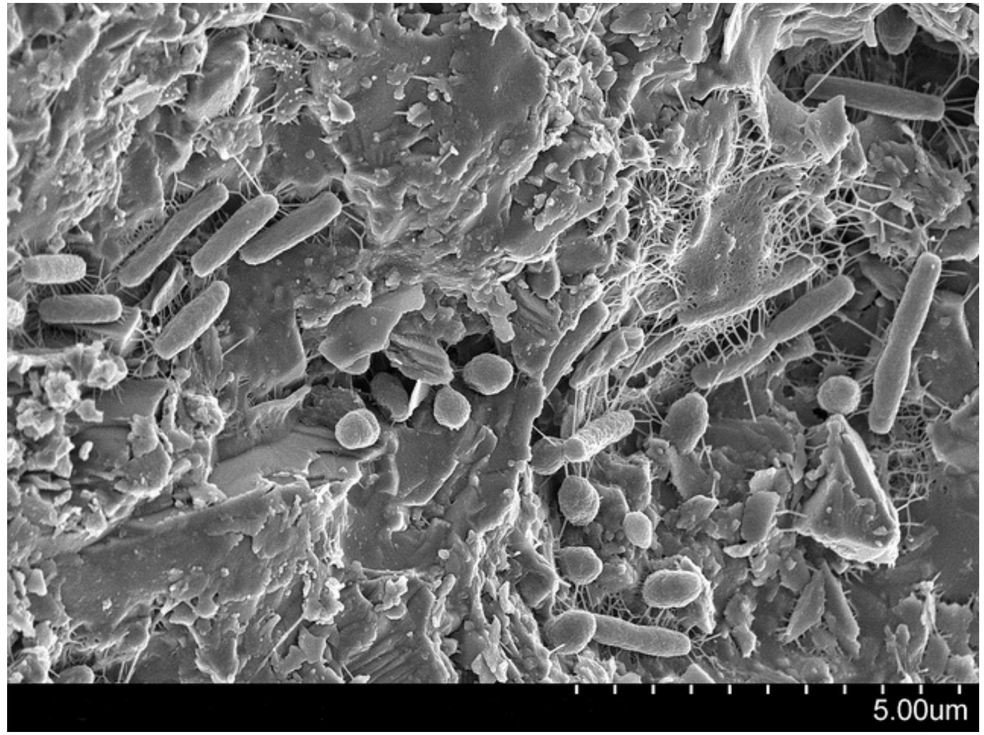


Metagenomics emerged in response to the observation that most micro-organisms can't be cultured

Morphological diversity typical of microorganisms cultured from soil on a broad spectrum medium, tryptic soy agar.

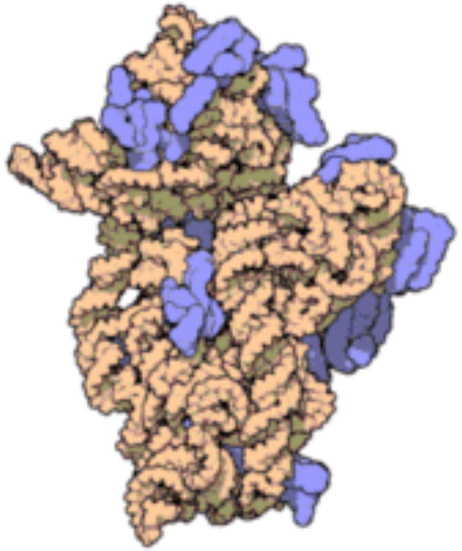
Handelsman et al (1998) Chem & Biol

Estimated: 10^8 bacteria per gram of soil of 6000-8000 different species
Only ~1% culturable (?)

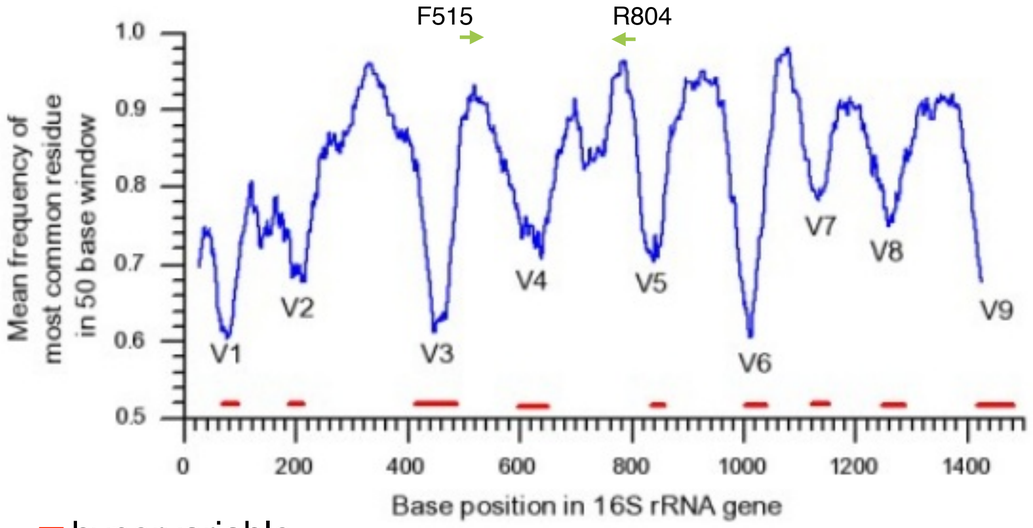


EM: Kim Lewis, Northeastern Univ.

PCR using primers targeting conserved regions of the 16S rRNA gene and sequencing enables genotyping of bacteria and archaea without having to culture them

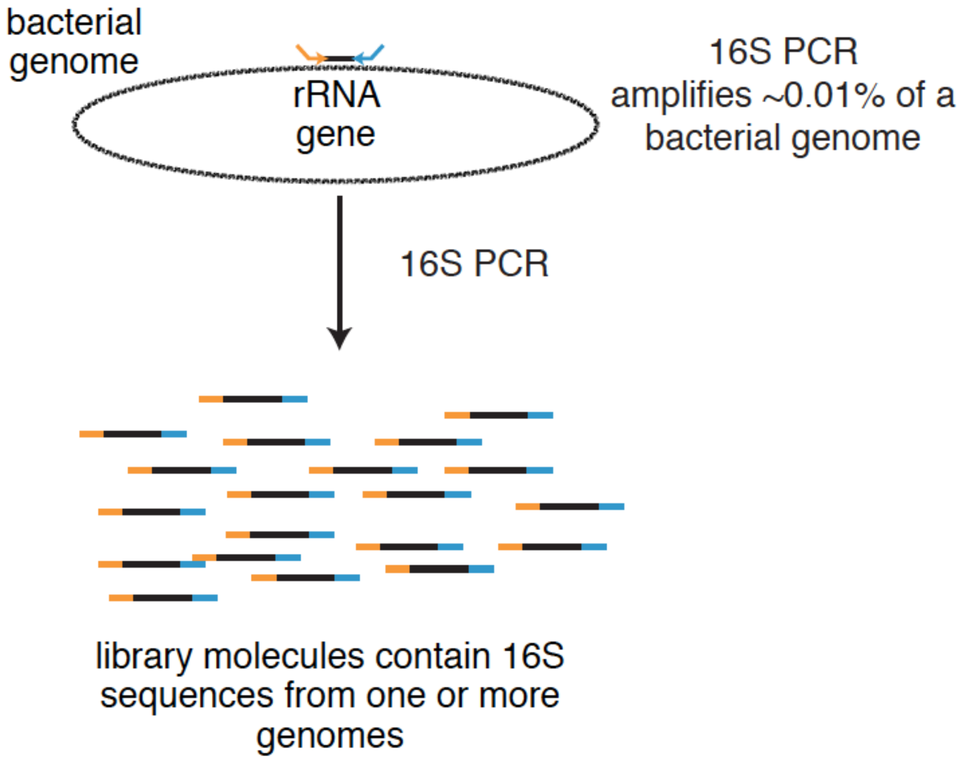


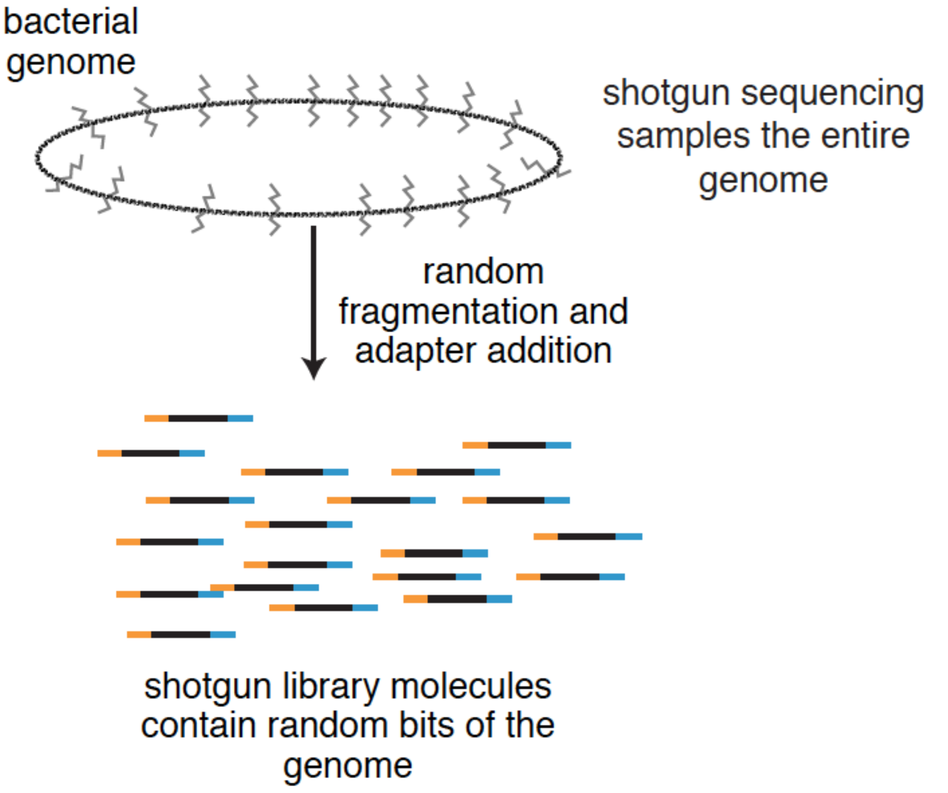
bacterial 30S ribosomal subunit
16S rRNA is in orange
(purple: ribosomal proteins)
image: wikipedia



hypervariable

16S sequencing vs. shotgun metagenomics



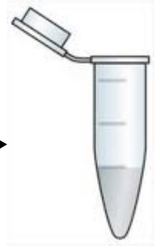


- Only bacteria and archaea surveyed
- Deeper sampling of bacterial diversity per $
- Relatively easy to make libraries and interpret results
- Appropriate if all you care about is microbial diversity / ecology

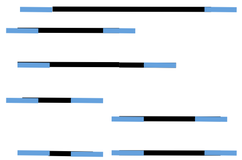
- All organisms studied*
- Decreased sampling depth per $
- Enables analysis of other genomic features of organisms, e.g. antimicrobial resistant genes
- Analysis is significantly more difficult

Pathogen discovery using metagenomic sequencing

case and control tissues

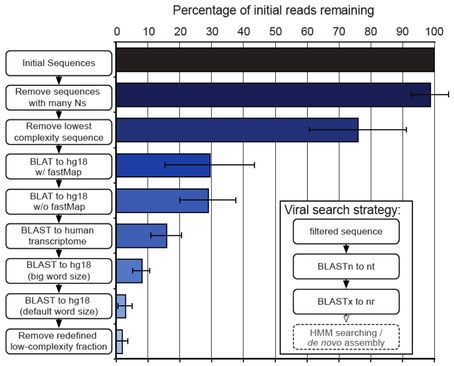


Nucleic acid



Library prep / barcode

Illumina sequencing

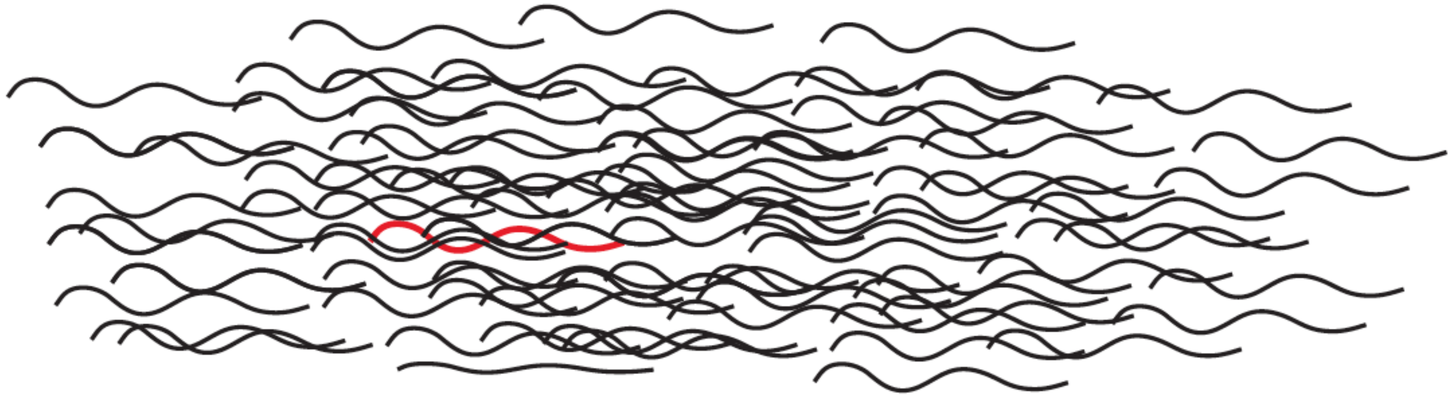


Computational Analysis

Follow-up

A couple of the key challenges in metagenomics pathogen discovery

1) Pathogen nucleic acid is typically present in a sea of host nucleic acid



~1 viral nucleic acid per 10^4-10^7 host nucleic acids

2) New pathogens have unknown sequences

TTTCAG?TTT?ACC????TG??AAA?ACATCC??TATAC???T?

3) Misannotated sequences in databases confound results

How sensitive is NGS for pathogen detection?
In theory, a single read is sufficient to identify a pathogen
(but that's cutting it a little close)

A single read pair aligning to **caprine herpesvirus-2**
amongst ~0.5M mule deer reads

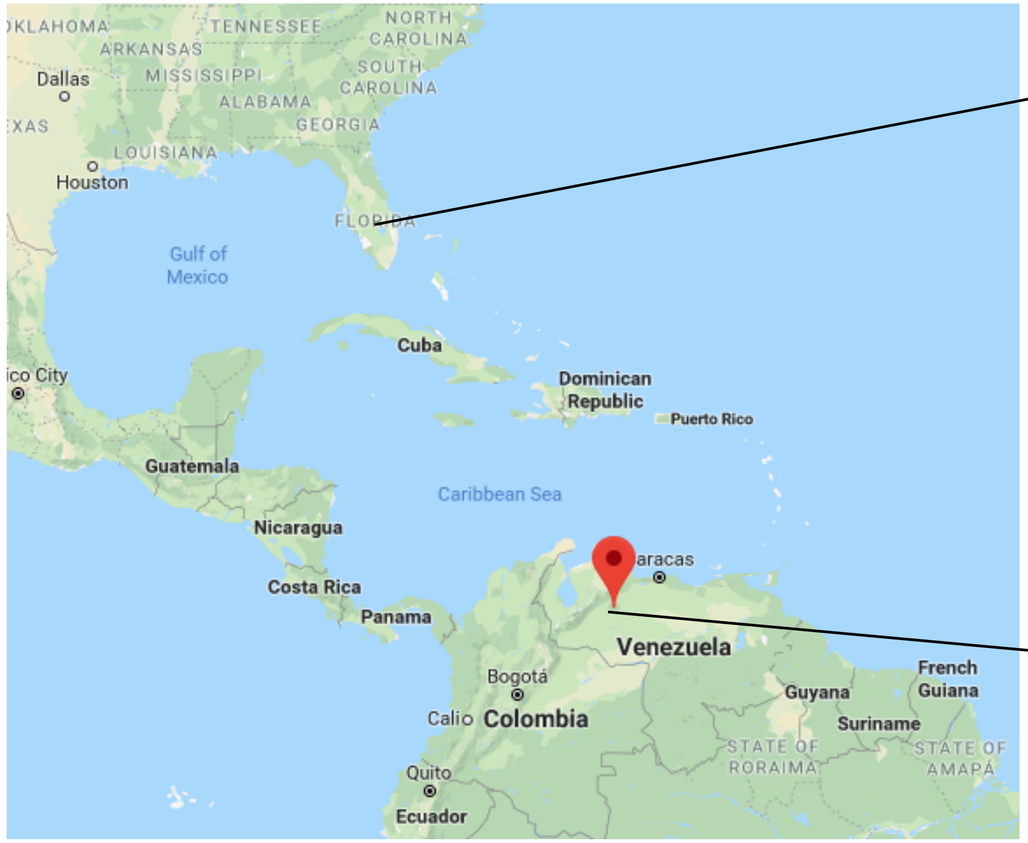
Identification of this pathogen completely consistent with histopathology

case had been tested for *ovine* herpesvirus 2

Laura Hoon-Hanks, DVM
Samples from: Karen Fox DVM, CO Parks & Wildlife

PCR is generally more sensitive than NGS for targeted pathogen detection

Example of problematic annotation: Guanarito virus sequence supposedly in a pool of *Culex cedecei* mosquitoes collected in the Florida Everglades



Arenavirus
Cause of Venezuelan hemorrhagic fever
Not known to be arthropod borne

image: American Society of Mammalogists

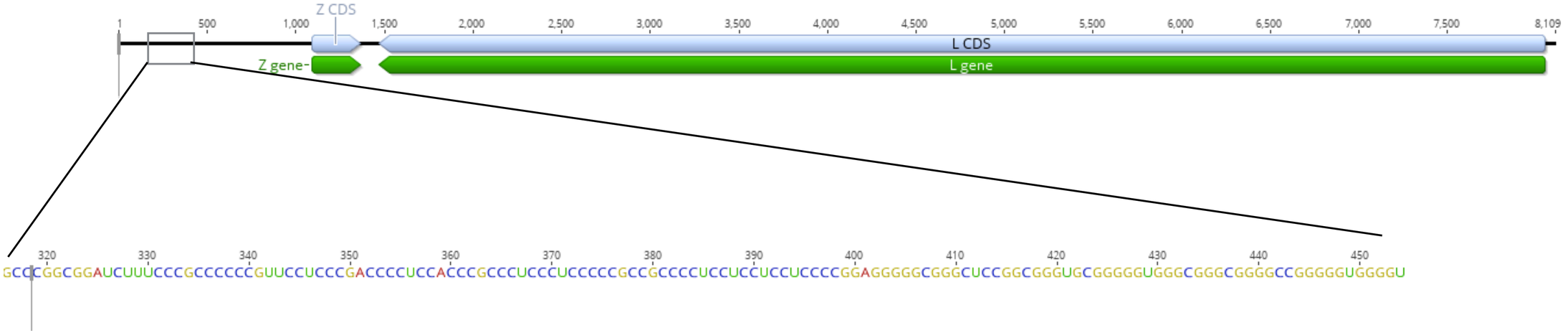
One of the putative Guanarito virus sequences

>NODE_274_length_640_cov_11681_ID_547
GCGGGGGTGGGCGGGCGGGGCCGGGGGTGGGGTCGGCGGGGGACCGTCCCCCGACCGGCGACCGGCCGCCGCCGGGC
GCATTTCCACCGCGGCGGTGCGCCGCGACCGGCTCCGGGACGGCTGGGAAGGCCCGGCGGGGAAGGTGGCTCGGGGG
GCCCCGTCCCGCCCCGTCTTCCCCCCGCCCGCGTCCTCCCCCGGGAGGGCGCGGGTCGGGGCGGCGGCGGCGGTGGC
GGCGGGACCACCCCCCGAGTGTTACAGCCCCCCGGCAGCAGCACTCGCCGAATCCCGGGGCCGAGGGAGCGAGACCC
GTCGCCGCGCTCTCCCCCCTCCCGGCGCCCACCCCCGCGGGGGCCCCCCGCGCGGGGGTCCCCCCCCGCGGGGGCGCG
CCGGCGGTCCTCGTGGGGGGCCGGGCCACCCCTCCCACGGCGCGACCGCTCTCCCACCCCCTCCCCGCACCCCCGGC
GACGGGGGCCCGCGCGGGTGGGGGCGGGGCGGACTGTCCCCAGTGCGCCCCGGGCGGGTCGCGCCGTCGGGCCCGGG
GGGGTTCTCTCGGGGCCACGCGCGCGTCCCTCGAAGAGGGGGACGGCGGAGCGAGCGCACGGGGTCGGCGGCGATGT
CGGCTACCCACCCGACCCGTCTTG

BLAST the sequence vs. the NCBI nucleotide database

Description	Max score	Total score	Query cover	E value	Ident	Accession
Guanarito mammarenavirus isolate CVH-960201 segment L, complete sequence	1081	1081	94%	0.0	99%	KU746283.1
Guanarito mammarenavirus isolate CVH-950801 segment S, complete sequence	1064	1064	90%	0.0	99%	KU746280.1
Chimpanzee 28S ribosomal RNA gene fragment	826	826	100%	0.0	89%	M30950.1
Gorilla 28S ribosomal RNA gene fragment	817	817	99%	0.0	89%	M30951.1
Homo sapiens external transcribed spacer 18S ribosomal RNA gene, internal transcribed spacer 1, 5.8S ribosomal RNA gene, internal transcribed spacer 2, 28S ribosomal RNA gene, and e	808	808	100%	0.0	89%	KY962518.1
Homo sapiens clone BAC JH1 genomic sequence	808	1612	100%	0.0	89%	MF164269.1
Homo sapiens RNA, 45S pre-ribosomal N2 (RNA45SN2), ribosomal RNA	804	804	100%	0.0	89%	NR_146144.1
Homo sapiens RNA, 28S ribosomal N2 (RNA28SN2), ribosomal RNA	804	804	100%	0.0	89%	NR_146148.1
Human DNA sequence from clone CH507-146P16 on chromosome 21, complete sequence	804	804	100%	0.0	89%	CT476837.18
Human ribosomal DNA complete repeating unit	798	798	100%	0.0	89%	U13369.1
Homo sapiens clone BAC JH5 genomic sequence	787	787	100%	0.0	88%	MF164266.1
Homo sapiens RNA, 28S ribosomal N3 (RNA28SN3), ribosomal RNA	784	784	100%	0.0	88%	NR_146154.1
Homo sapiens RNA, 45S pre-ribosomal N3 (RNA45SN3), ribosomal RNA	784	784	100%	0.0	88%	NR_146151.1
Human DNA sequence from clone CH507-528H12 on chromosome 21, complete sequence	784	1707	100%	0.0	88%	FP236383.15

These Guanarito virus sequences are mis-assembled



Conclusion: don’t blindly trust database annotation nor the output of analysis software

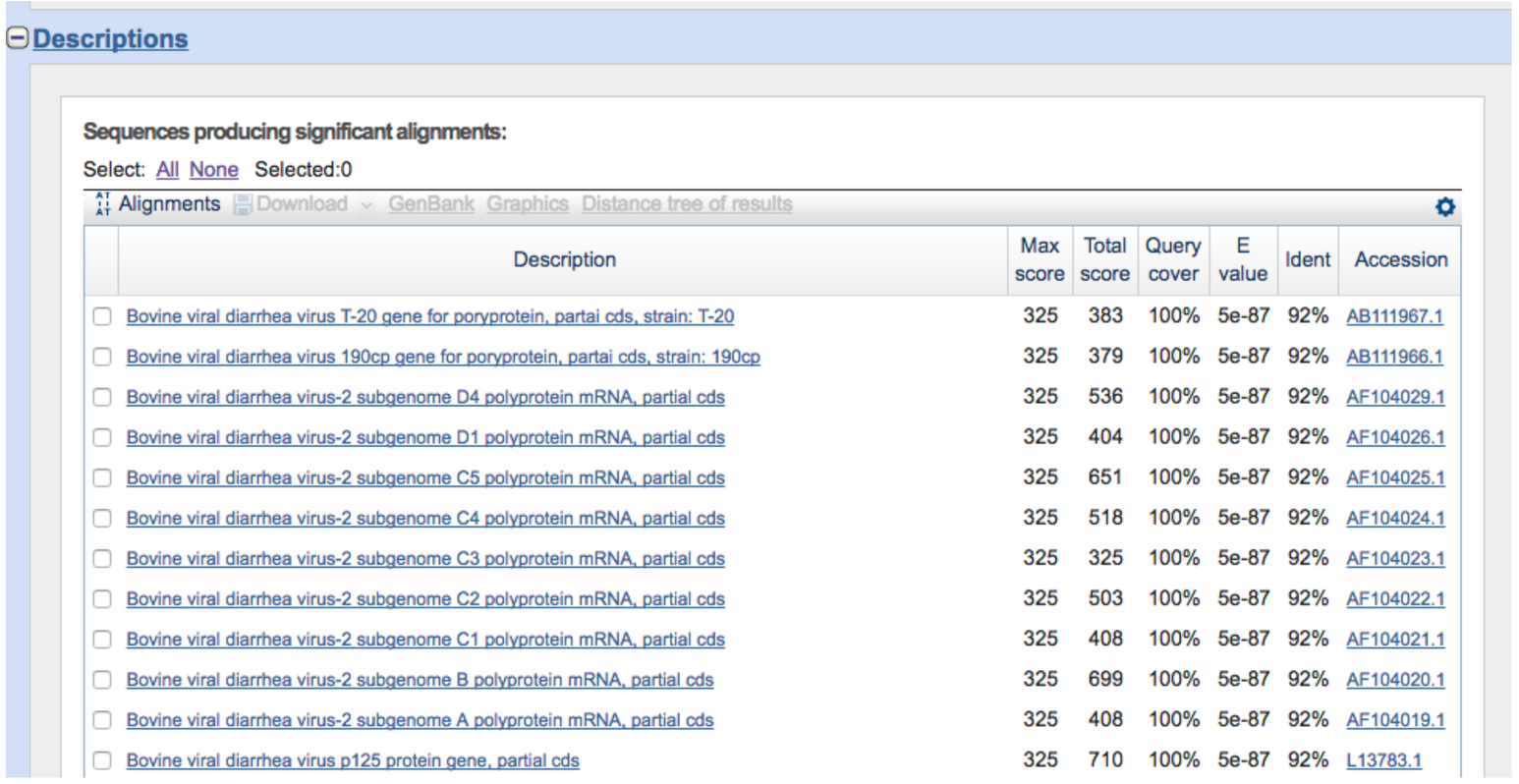
```
COMMENT     GenBank Accession Numbers KU746283, KU746284 represent sequences
            from the 2 segments of Guanarito mammarenavirus CVH-960201.

            ##Genome-Assembly-Data-START##
            Assembly Method :: Trimmomatic v. 0.32
            SGA v. 0.10.13
            IMETAMOS v. 1.5
            samtools v1.1
            FastQC v. 0.10.0
            Spades v. 3.1.1
            idba v1.1.1
            Pilon v. 1.8
            Quast v. 2.2
            Prokka v. 1.7
            Assembly Name :: GTOV014-SEQ-1-ASM-1
            Genome Coverage :: 6779.96x
            Sequencing Technology :: Illumina HiSeq1500
            ##Genome-Assembly-Data-END##.
FEATURES             Location/Qualifiers
     source          1..8109
                     /organism="Guanarito mammarenavirus"
                     /mol_type="genomic RNA"
```

Another caveat: using smaller databases (e.g. all viral genomes in RefSeq) is faster but it can produce misleading results

Here: a read was BLASTed against all of the virus nucleotide sequences in Genbank

```
>a_sequence
ATGCAGATCTTCGTGAAGACTCTGACTGGTAAGACCATCACCCTCGAGGTTGAGCC...
```



Descriptions

Sequences producing significant alignments:

Select: All None Selected:0

Alignments Download GenBank Graphics Distance tree of results

Description	Max score	Total score	Query cover	E value	Ident	Accession
Bovine viral diarrhea virus T-20 gene for poryprotein, partai cds, strain: T-20	325	383	100%	5e-87	92%	AB111967.1
Bovine viral diarrhea virus 190cp gene for poryprotein, partai cds, strain: 190cp	325	379	100%	5e-87	92%	AB111966.1
Bovine viral diarrhea virus-2 subgenome D4 polyprotein mRNA, partial cds	325	536	100%	5e-87	92%	AF104029.1
Bovine viral diarrhea virus-2 subgenome D1 polyprotein mRNA, partial cds	325	404	100%	5e-87	92%	AF104026.1
Bovine viral diarrhea virus-2 subgenome C5 polyprotein mRNA, partial cds	325	651	100%	5e-87	92%	AF104025.1
Bovine viral diarrhea virus-2 subgenome C4 polyprotein mRNA, partial cds	325	518	100%	5e-87	92%	AF104024.1
Bovine viral diarrhea virus-2 subgenome C3 polyprotein mRNA, partial cds	325	325	100%	5e-87	92%	AF104023.1
Bovine viral diarrhea virus-2 subgenome C2 polyprotein mRNA, partial cds	325	503	100%	5e-87	92%	AF104022.1
Bovine viral diarrhea virus-2 subgenome C1 polyprotein mRNA, partial cds	325	408	100%	5e-87	92%	AF104021.1
Bovine viral diarrhea virus-2 subgenome B polyprotein mRNA, partial cds	325	699	100%	5e-87	92%	AF104020.1
Bovine viral diarrhea virus-2 subgenome A polyprotein mRNA, partial cds	325	408	100%	5e-87	92%	AF104019.1
Bovine viral diarrhea virus p125 protein gene, partial cds	325	710	100%	5e-87	92%	L13783.1

Cool, looks like a flavivirus! Right?

Keep analyses as unbiased as possible

The same read was BLASTed against all the nucleotide sequences in Genbank (the ‘nt’ database):

```
>a_sequence
ATGCAGATCTTCGTGAAGACTCTGACTGGTAAGACCATCACCCTCGAGGTTGAGCC...
```

Sequences producing significant alignments:

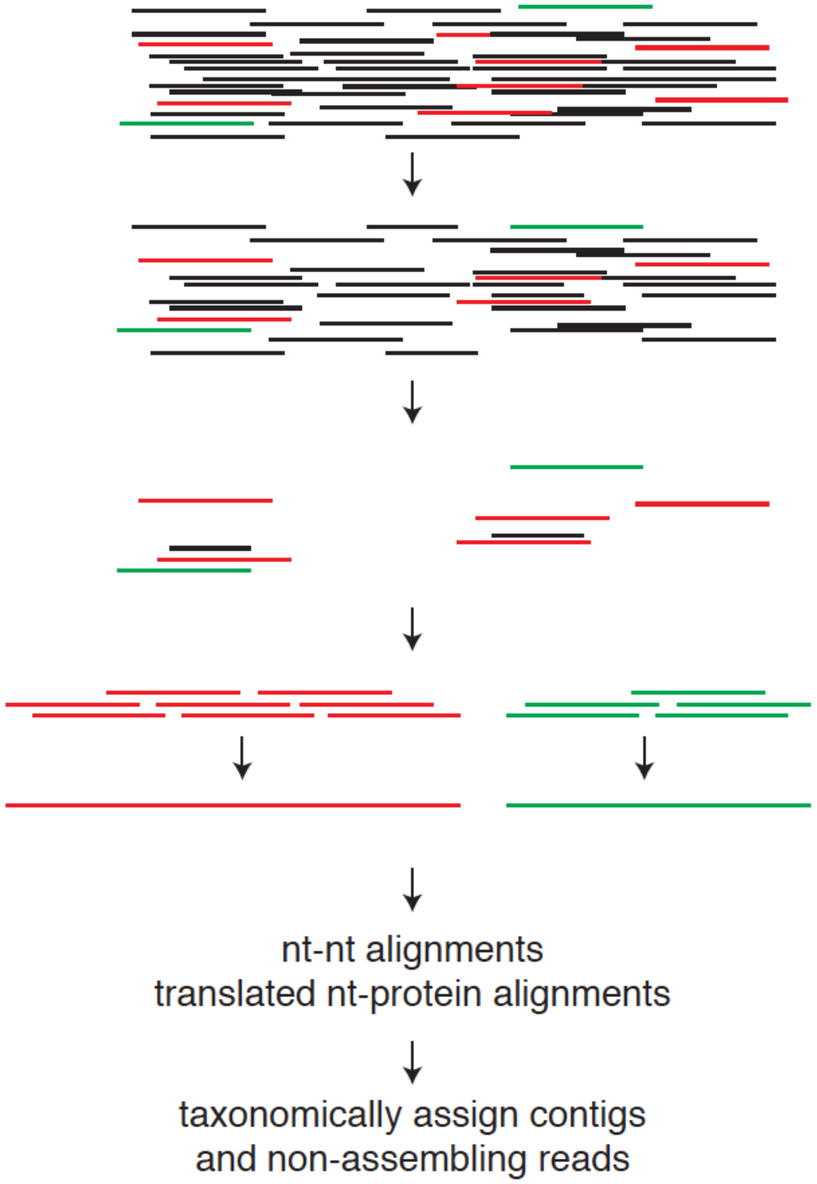
Select: All None Selected:0

Alignments Download GenBank Graphics Distance tree of results

Description	Max score	Total score	Query cover	E value	Ident	Accession
Homo sapiens ubiquitin C (UBC), RefSeqGene on chromosome 12	412	3261	100%	2e-111	100%	NG_027722.2
Homo sapiens ubiquitin C (UBC), mRNA	412	3261	100%	2e-111	100%	NM_021009.6
Homo sapiens UbC gene for ubiquitin C, complete cds, clone: MKJ9	412	3241	100%	2e-111	100%	AB643790.1
Homo sapiens UbC gene for ubiquitin C, complete cds, clone: DJL8	412	2881	100%	2e-111	100%	AB643789.1
Homo sapiens UbC gene for ubiquitin C, complete cds, clone: DJL9	412	3266	100%	2e-111	100%	AB643788.1
Homo sapiens UbC gene for ubiquitin C, complete cds, clone: MKS7	412	2558	100%	2e-111	100%	AB643787.1
Homo sapiens UbC gene for ubiquitin C, complete cds, clone: MKS9	412	3257	100%	2e-111	100%	AB643786.1
Homo sapiens UbC gene for ubiquitin C, complete cds, clone: BHP7	412	2549	100%	2e-111	100%	AB643785.1
Pan troglodytes mRNA for ubiquitin, complete cds, clone: PtsC-51-5_D12	412	1833	100%	2e-111	100%	AK306071.1

Some BVDV genomes contain ubiquitin homologs

A typical pathogen discovery analysis workflow



Quality filter /
remove PCR duplicates
cutadapt
cd-hit-est

Remove host sequences
bowtie2

Assemble remaining reads into contigs
SPAdes

nt-nt alignments
translated nt-protein alignments

taxonomically assign contigs and non-assembling reads

Identify most similar sequences in NCBI databases
BLASTN (gsnap),
BLASTX (diamond)

~40 min for a dataset w/ 6M reads

caveat: for a dataset where host filtering removed almost all of the reads

A good reference genome is helpful but not strictly necessary

Ixodes scapularis
Blacklegged tick

genome sequenced
(Gulia-Nuss et al (2015) Nature Comm)

Images: CDC

~3% of reads remaining after filtering

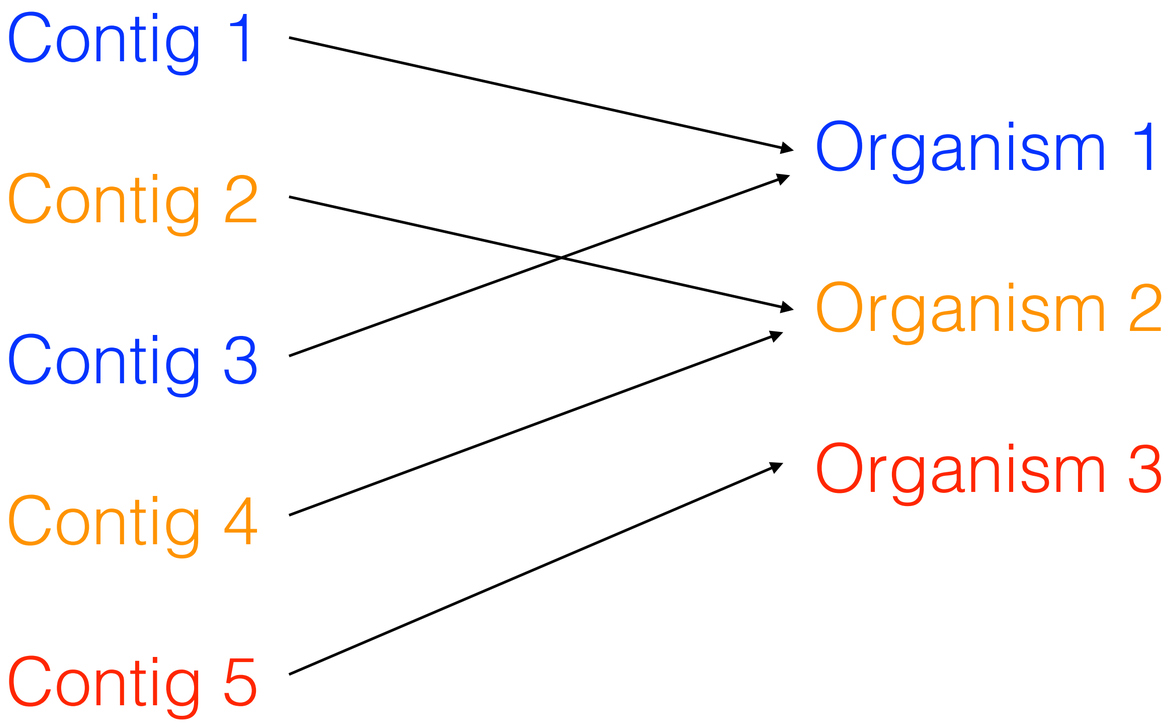
Dermacentor andersoni
Rocky mountain wood tick

no genome sequence

~55% of reads remaining after filtering

- Assembly/downstream steps will go much slower
- The taxonomic assessment will be more difficult (lots of false positive taxonomic assignments)

The goal of metagenomic classification software is to map sequence information to taxonomic information.



Earlier, we did this by BLASTing several contigs on the NCBI website. This is not a practical approach for many contigs.

Nucleotide-level similarity identifies closely related organisms (blastn-like (blastn))
Protein-level similarity discovers ‘new’ organisms (blastx-like (diamond))

A nice review of metagenomic classifiers

Briefings in Bioinformatics, 2017, 1–15

doi: 10.1093/bib/bbx120
Paper

A review of methods and databases for metagenomic classification and assembly

Florian P. Breitwieser, Jennifer Lu and Steven L. Salzberg

Corresponding author: Steven L. Salzberg, Center for Computational Biology, Johns Hopkins University, 1900 E. Monument St., Baltimore, MD, 21205, USA.
E-mail: salzberg@jhu.edu

Name	References	URL
CaPSID	Borozan et al., 2012	https://github.com/capsid/capsid
ClassyFlu	Van der Auwera et al., 2014	http://bioinf.uni-greifswald.de/ClassyFlu/query/init
Clinical PathoScope	Byrd et al., 2014	https://sourceforge.net/p/pathoscope/wiki/clinical_pathoscope/
DUDes	Piro et al., 2016	http://sf.net/p/dudes
EnsembleAssembler	Deng et al., 2015	https://github.com/xutaodeng/EnsembleAssembler
Exhaustive Iterative Assembly (Virus Discovery Pipeline)	Schürch et al., 2014	–
FACS	Stranneheim et al., 2010	https://github.com/SciLifeLab/facs
GenSeed-HMM	Alves et al., 2016	https://sourceforge.net/projects/genseedhmm/
Giant Virus Finder	Kerepesi and Grolmusz, 2016	http://pitgroup.org/giant-virus-finder
GOTTCHA	Freitas et al., 2015	https://github.com/LANL-Bioinformatics/GOTTCHA
IMSA	Dimon et al., 2013	https://sourceforge.net/projects/arron-imsa/?source=directory
IMSA+A	Cox et al., 2017	https://github.com/JeremyCoxBMI/IMSA-A
Kraken	Wood and Salzberg, 2014	https://github.com/DerrickWood/kraken
LMAT	Ames et al., 2013	https://sourceforge.net/projects/lmat/
MEGAN 4	Huson et al., 2011	http://ab.inf.uni-tuebingen.de/software/megan4/
MEGAN Community Edition	Huson et al., 2016	http://ab.inf.uni-tuebingen.de/data/software/megan6/download/welcome.html
MePIC	Takeuchi et al., 2014	https://mepic.nih.go.jp/
MetaShot	Fosso et al., 2017	https://github.com/bfosso/MetaShot
metaVIC	Modha, 2016	https://github.com/sejmodha/metaVIC
Metavir	Roux et al., 2011	http://metavir-meb.univ-bpclermont.fr/
Metavir 2	Roux et al., 2014	http://metavir-meb.univ-bpclermont.fr/
MetLab	Norling et al., 2016	https://github.com/norling/metlab
NBC	Rosen et al., 2011	http://nbc.ece.drexel.edu/
PathSeq	Kostic et al., 2011	https://www.broadinstitute.org/software/pathseq/
ProViDE	Ghosh et al., 2011	http://metagenomics.atc.tcs.com/binning/ProViDE/
QuasQ	Poh et al., 2013	http://www.statgen.nus.edu.sg/$\sim$software/quasq.html
READSCAN	Naeem et al., 2013	http://cbrc.kaust.edu.sa/readscan/
Rega Typing Tool	Kroneman et al., 2011; Pineda-Peña et al., 2013	http://regatools.med.kuleuven.be/typing/v3/hiv/typingtool/
RIEMS	Scheuch et al., 2015	https://www.fli.de/fileadmin/FLI/IVD/Microarray-Diagnostics/RIEMS.tar.gz
RINS	Bhaduri et al., 2012	http://khavarilab.stanford.edu/tools-1/#tools
SLIM	Cotten et al., 2014	"Available upon request"
SMART	Lee et al., 2016	https://bitbucket.org/ayl/smart
SRSA	Isakov et al., 2011	"Available upon request"
SURPI	Naccache et al., 2014	https://github.com/chiulab/surpi
Taxonomer	Flygare et al., 2016	https://www.taxonomer.com/
Taxy-Pro	Klingenberg et al., 2013	http://gobics.de/TaxyPro/
"Unknown pathogens from mixed clinical samples"	Gong et al., 2016	–
vFam	Skewes-Cox et al., 2014	http://derisilab.ucsf.edu/software/vFam/
VIP	Li et al., 2016	https://github.com/keylabivdc/VIP
ViralFusionSeq	Li et al., 2013	https://sourceforge.net/projects/viralfusionseq/
Virana	Schelhorn et al., 2013	https://github.com/schelhorn/virana
VirFind	Ho and Tzanetakis, 2014	http://virfind.org/j/
VIROME	Wommack et al., 2012	http://virome.dbi.udel.edu/app/#view=home
ViromeScan	Rampelli et al., 2016	https://sourceforge.net/projects/viromescan/
VirSorter	Roux et al., 2015	https://github.com/simroux/VirSorter
VirusFinder	Wang et al., 2013	http://bioinfo.mc.vanderbilt.edu/VirusFinder/
VirusHunter	Zhao et al., 2013	https://www.ibridgenetwork.org/#!/profiles/9055559575893/innovations/103/
VirusSeeker	Zhao et al., 2017	https://wupathlabs.wustl.edu/virusseeker/
VirusSeq	Chen et al., 2013	http://odin.mdacc.tmc.edu/$\sim$xsu1/VirusSeq.html
VirVarSeq	Verbist et al., 2015	http://sourceforge.net/projects/virtools/?source=directory
VMGAP	Lorenzi et al., 2011	–

–, *No website could be found, the workflow was unavailable.*

Metagenomic classification can be challenging

Resource intensive
Large databases
Large assemblies
Memory and storage intensive

Bioinformatics challenges

User-friendly bioinformatics software for analysis of mNGS data is not currently available. Thus, customized bioinformatics pipelines for analysis of clinical mNGS data[56,109–111] still require highly trained programming staff to develop, validate and maintain the pipeline for clinical use. The laboratory can either host computational servers locally or move the bioinformatics analysis and data storage to cloud platforms. In either case, hardware and software setups can be complex, and adequate measures

Clinical metagenomics

Charles Y. Chiu[1,2]* and Steven A. Miller[1]

350 | JUNE 2019 | VOLUME 20

Nat Rev Genetics

idseq.net is a new web-based tool that does metagenomic classification

it's goal is to provide user-friendly metagenomic classification

CHAN ZUCKERBERG BIOHUB

IDseq is an unbiased global software platform that helps scientists identify pathogens in metagenomic sequencing data.

Discover
Identify the pathogen landscape

Detect
Monitor and review potential outbreaks

Decipher
Find potential infecting organisms in large datasets

Learn more about IDseq

Already have an account? Sign in.

First Name **Last Name**

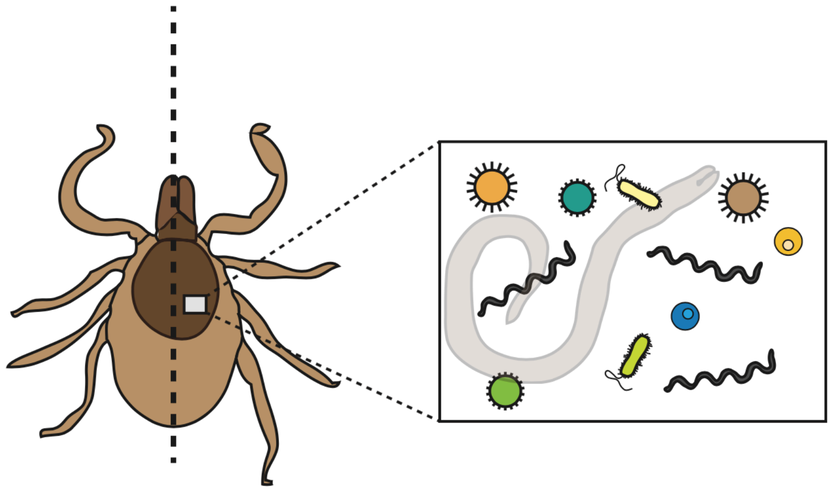
Email

Affiliated Institution or Company

How would you use IDseq? Optional

Submit

Name	Total reads	Passed filters	Passed QC	DCR	Host	Collection Location	Status
Tick_16 a year ago | Mark Stenglein	**14,822,970**	**377,104** 2.54%	**61.82%**	**1.75**	**Tick**	--	✔ COMPLETE 50 minutes
Tick_15 a year ago | Mark Stenglein	**8,776,796**	**202,590** 2.31%	**60.93%**	**1.51**	**Tick**	--	✔ COMPLETE 47 minutes
Tick_14 a year ago | Mark Stenglein	**14,132,738**	**307,554** 2.18%	**63.27%**	**1.48**	**Tick**	--	✔ COMPLETE 46 minutes
Tick_13 a year ago | Mark Stenglein	**14,892,862**	**356,202** 2.39%	**59.85%**	**1.58**	**Tick**		
Tick_12 a year ago | Mark Stenglein	**13,067,958**	**199,586** 1.53%	**60.72%**	**1.38**			
Tick_11 a year ago | Mark Stenglein	**14,281,564**	**449,058** 3.14%	**63.74%**	**1.36**			
Tick_10 a year ago | Mark Stenglein	**11,685,018**	**200,632** 1.72%	**61.54%**	**1.29**			



PIPELINE v1.7, NT/NR: 2018-02-15 | processed a year ago

Stenglein_I_scap_ticks >

Tick_16

Sample Details

Share Download

Taxon name | Name Type: Scientific | Background: NID Human CSF v3 | Categories | Threshold Filters | Read Specificity: All | Min Contig Size: 4

999 rows passing filters, out of 999 total rows.

Taxon	Score	Z	rPM	r	%id	L	log(1/E)	NT / NR
Phlebovirus *(3 viral species)* ● 3 PATHOGENIC | A	**36,880,366**	**100.0** 99.0	**3,667.1** 34.5	**54,358** 512	**98.6** 100.0	**72.2** 24.4	**36** 7	
Non-genus-specific reads in family Enterobacteriaceae *(1 bacterial species)*	**36,794,662**	**99.0** 45.9	**3,702.1** 31.4	**54,876** 465	**99.9** 99.9	**73.6** 24.4	**38** 7	
Ehrlichia *(15 bacterial species)* ● 15 PATHOGENIC | C	**22,546,226**	**99.0** 66.6	**3,838.3** 2.8	**56,895** 42	**99.1** 100.0	**71.3** 24.6	**36** 6	
Ixodes *(11 eukaryotic species)*	**15,789,538**	**99.0** -0.3	**2,005.1** 0.1	**29,722** 2	**98.1** 100.0	**69.5** 25.0	**34** 6	
Borreliella *(12 bacterial species)* ● 1	**3,867,729**	**99.0** 1.3	**731.3** 0.2	**10,840** 3	**99.9** 100.0	**73.1** 24.7	**38** 6	
Non-genus-specific reads in family Borreliaceae *(1 bacterial species)*	**1,692,640**	**100.0** -100.0	**169.3** 0.0	**2,509** 0	**99.9** 0.0	**71.2** 0.0	**37** 0	
Non-genus-specific reads in family Anaplasmataceae *(1 bacterial*	**323,922**	**99.0**	**32.6**	**483**	**99.9**	**51.9**	**24**	

PIPELINE v1.7, NT/NR: 2018-02-15 | processed a year ago

Stenglein_I_scap_ticks >

Tick_16

Sample Details

Share | Download

- Download Report Table (.csv)
- Download Non-Host Reads (.fasta)
- Download Unmapped Reads (.fasta)
- See Results Folder
- Download Taxon Tree as SVG
- Download Taxon Tree as PNG

Taxon name | Name Type: Scientific | Background: NID Human CSF v3 | Categories | Threshold Filters | Read

cifici

999 rows passing filters, out of 999 total rows.

Taxon	Score	Z	rPM	r			
Phlebovirus *(3 viral species)* ● 3 PATHOGENIC | A	36,880,366	100.0 99.0	3,667.1 34.5	54,358 512	100.0	24.4	7
Non-genus-specific reads in family Enterobacteriaceae (1 bacterial species)	36,794,662	99.0 45.9	3,702.1 31.4	54,876 465	99.9 99.9	73.6 24.4	38 7
Ehrlichia *(15 bacterial species)* ● 15 PATHOGENIC | C	22,546,226	99.0 66.6	3,838.3 2.8	56,895 42	99.1 100.0	71.3 24.6	36 6
Ixodes *(11 eukaryotic species)*	15,789,538	99.0 -0.3	2,005.1 0.1	29,722 2	98.1 100.0	69.5 25.0	34 6
Borreliella *(12 bacterial species)* ● 1	3,867,729	99.0 1.3	731.3 0.2	10,840 3	99.9 100.0	73.1 24.7	38 6
Non-genus-specific reads in family Borreliaceae (1 bacterial species)	1,692,640	100.0 -100.0	169.3 0.0	2,509 0	99.9 0.0	71.2 0.0	37 0
Non-genus-specific reads in family Anaplasmataceae (1 bacterial	323,922	99.0	32.6	483	99.9	51.9	24

If you are interested in using idseq, contact

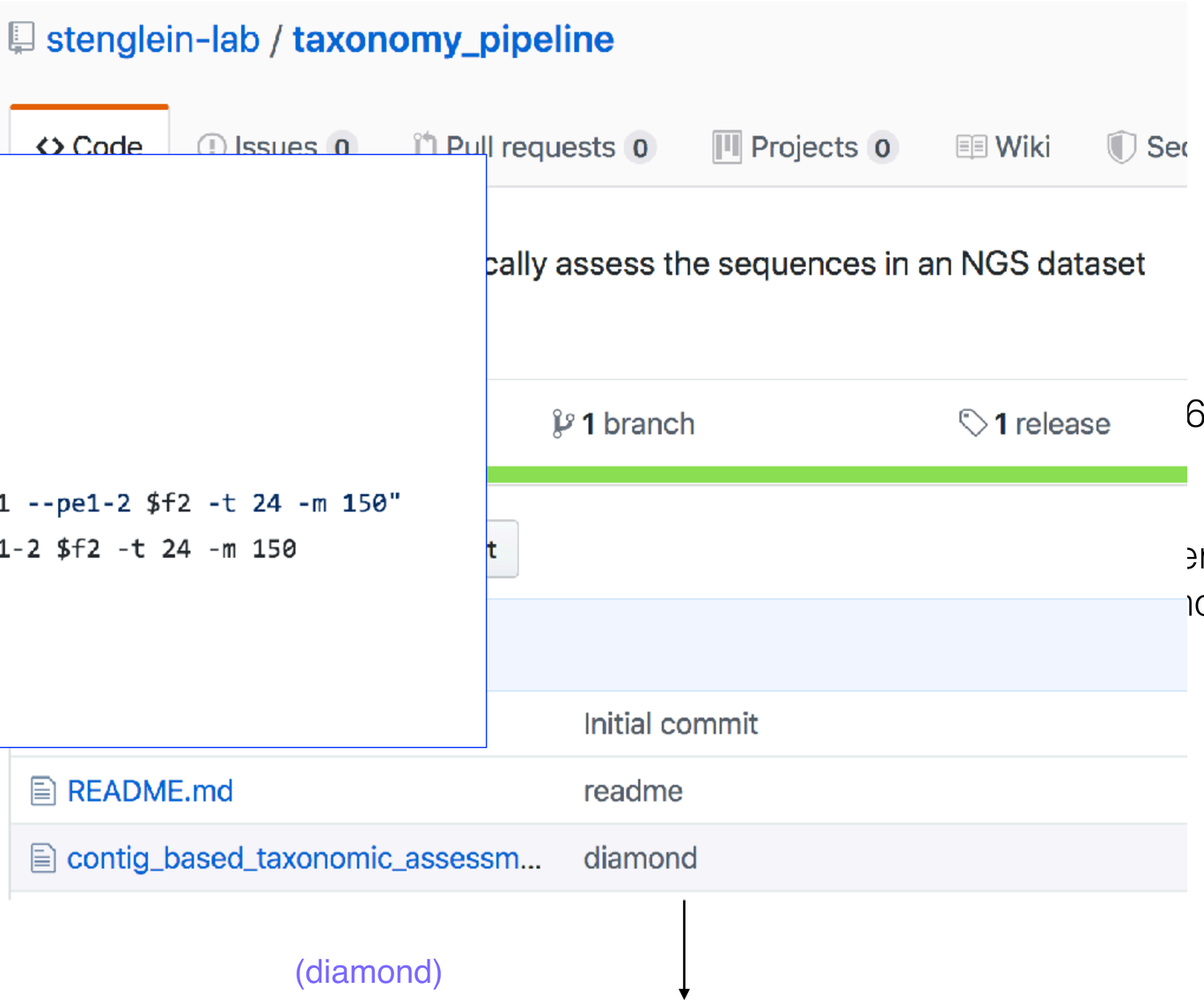
Rebecca Egger <regger@chanzuckerberg.com>

You may need to request that “your” reference genome gets added to the website

Our lab's pipeline is available on GitHub if you want to see how we do it.

```
61      # ***********************************
62      # first, create contigs using spades
63      # ***********************************
64
65      echo "run spades for $file_base"
66      date
67
68      echo "spades.py -o ${file_base}.spades --pe1-1 $f1 --pe1-2 $f2 -t 24 -m 150"
69      spades.py -o ${file_base}.spades --pe1-1 $f1 --pe1-2 $f2 -t 24 -m 150
70
71      echo "done running spades for $file_base"
72      date
73
```

- This pipeline is not easily portable!
- But at least you can see how we do it, which may be helpful.
- Using 'pipelines' facilitates reproducibility and throughput



https://github.com/stenglein-lab/taxonomy_pipeline

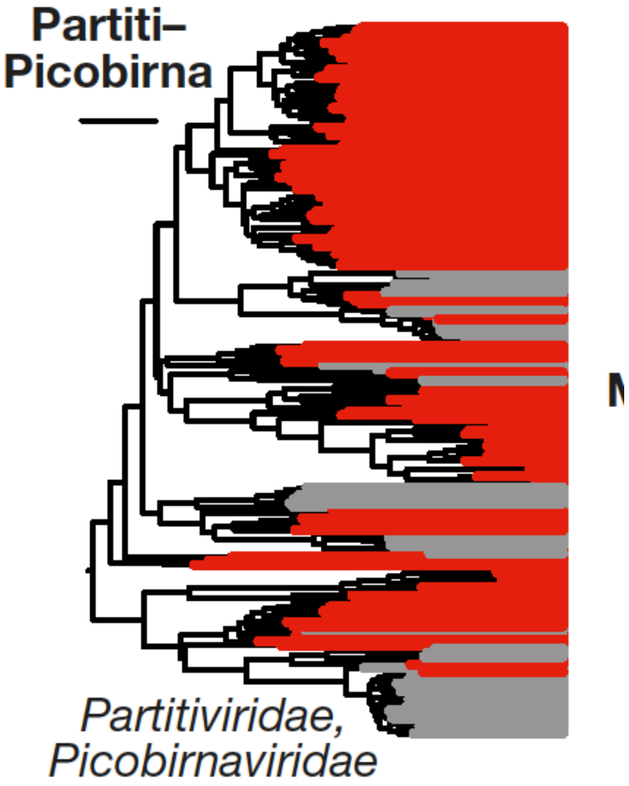
Metagenomic surveys of wild-caught organisms is leading to an explosion in discovery of new viral genome sequences

Redefining the invertebrate RNA virosphere

Mang Shi[1,2]*, Xian-Dan Lin[3]*, Jun-Hua Tian[4]*, Liang-Jun Chen[1]*, Xiao Chen[5]*, Ci-Xiu Li[1]*, Xin-Cheng Qin[1], Jun Li[6], Jian-Ping Cao[7], John Sebastian Eden[2], Jan Buchmann[2], Wen Wang[1], Jianguo Xu[1], Edward C. Holmes[1,2] & Yong-Zhen Zhang[1]

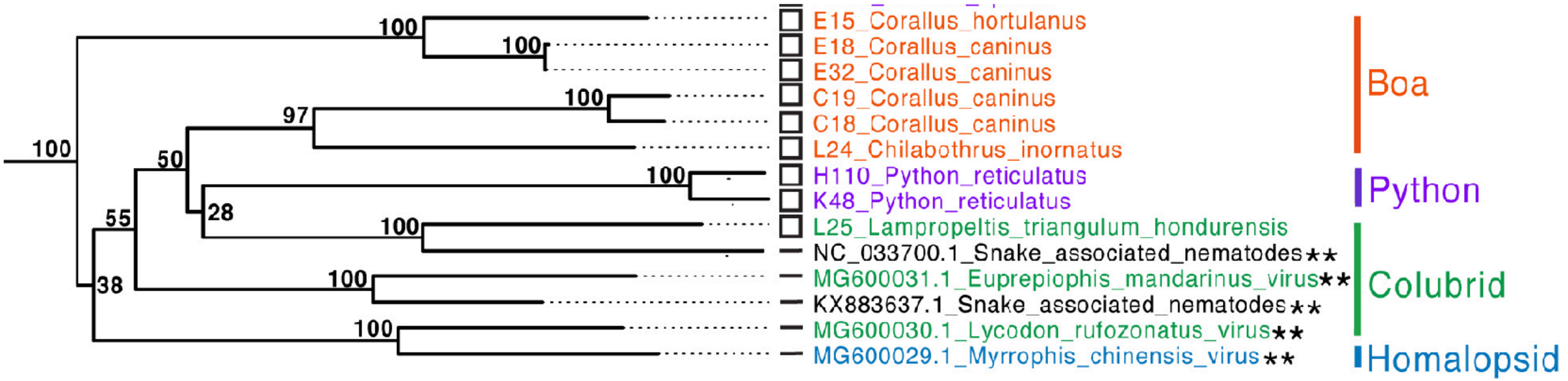
Current knowledge of RNA virus biodiversity is both biased and fragmentary, reflecting a focus on culturable or disease-causing agents. Here we profile the transcriptomes of over 220 invertebrate species sampled across nine animal phyla and report the discovery of 1,445 RNA viruses, including some that are sufficiently divergent to comprise new families. The identified viruses fill major gaps in the RNA virus phylogeny and reveal an evolutionary history that is characterized by both host switching and co-divergence. The invertebrate virome also reveals remarkable genomic flexibility that includes frequent recombination, lateral gene transfer among viruses and hosts, gene gain and loss, and complex genomic rearrangements. Together, these data present a view of the RNA virosphere that is more phylogenetically and genomically diverse than that depicted in current classification schemes and provide a more solid foundation for studies in virus ecology and evolution.

Nature **540**, 539–543 (22 December 2016)



Metagenomic sequencing only gives you sequences

Serpentoviruses detected in snakes with respiratory disease and also snake-associated nematodes

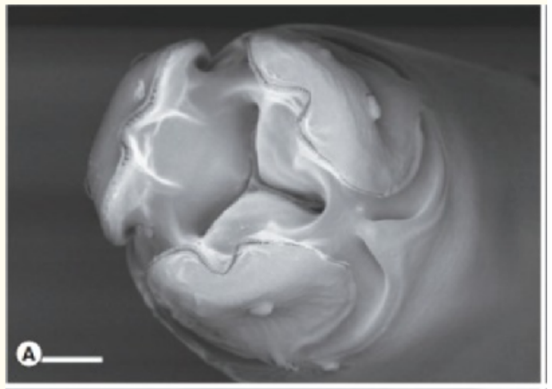


Are these related viruses infecting both nematodes and snakes?

I’d bet that the ‘nematode’ viruses really infect snakes

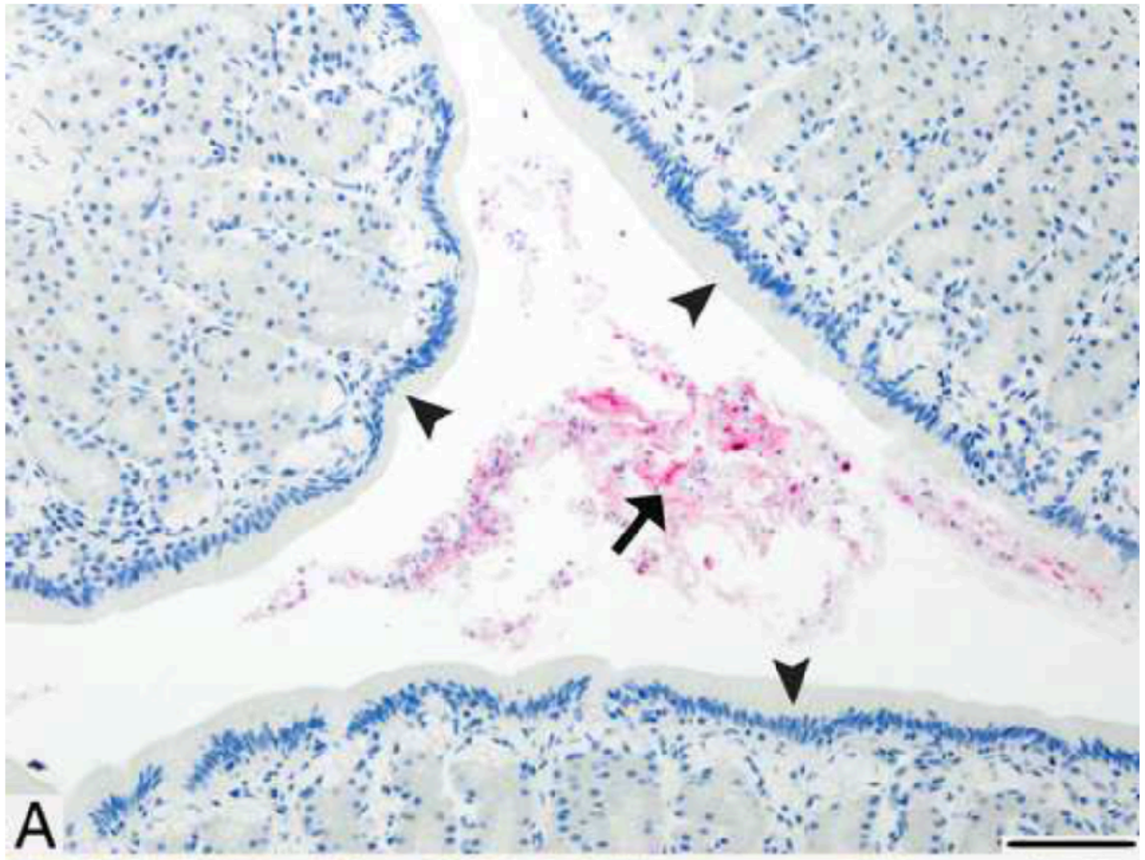
Laura Hoon-Hanks

SEM of a snake nematode



Choe et al (2016)

Serpentovirus antigen detected in python intestinal lumen



Sequencing can only give you sequences

A rabbit facility in TN experienced an outbreak of fatal gastroenteritis

image: wikipedia

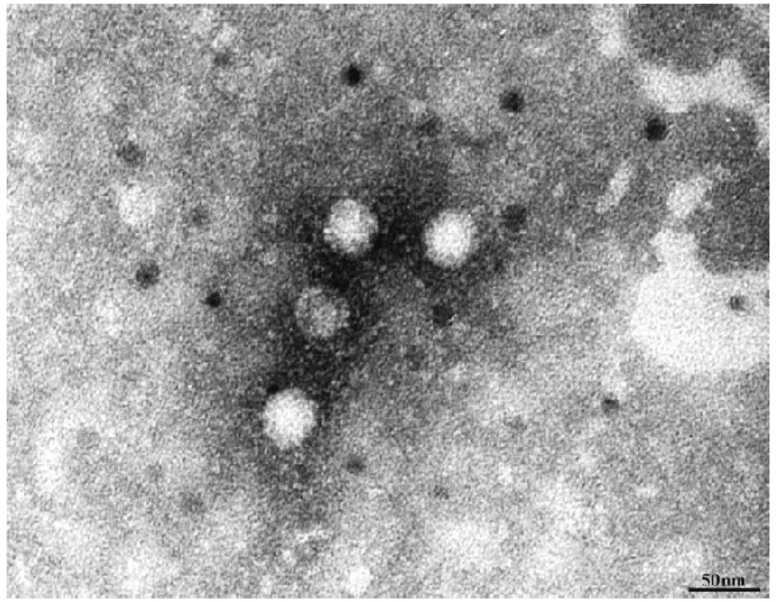


Figure 1 Electron micrograph of virus like particles in the stool of one animal (Table 1). Scale bar indicates 50 nm.

astrovirus sequences in the stool samples from sick rabbits

the virus is probably the cause of disease, but not proof

(Meta)genomics is useful for hypothesis generation but experiments must be done

Astrovirus particles

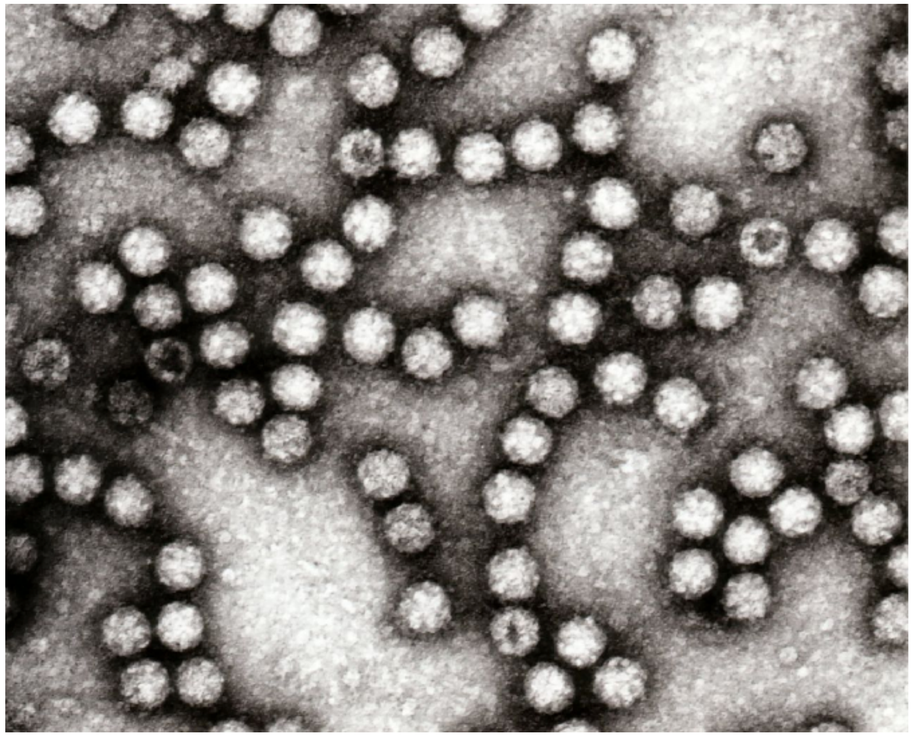


image: wikipedia

JOURNAL OF CLINICAL MICROBIOLOGY, Apr. 1993, p. 955–962
0095-1137/93/040955-08$02.00/0

Vol. 31, No. 4

Characterization and Seroepidemiology of a Type 5 Astrovirus Associated with an Outbreak of Gastroenteritis in Marin County, California

KAREN MIDTHUN,[1]†* HARRY B. GREENBERG,[1]‡ JOHN B. KURTZ,[2] G. WILLIAM GARY,[3] FENG-YING C. LIN,[4] AND ALBERT Z. KAPIKIAN[1]

RESULTS

Volunteer study. Nineteen adult volunteers were orally administered a filtrate prepared from a 0.1% suspension of stool from one of the ill individuals in the original Marin County outbreak. None of 17 volunteers who received a 1-ml inoculum became ill. Because of this, the amount of inoculum was increased to 20 ml. Of two volunteers who received the larger inoculum, one developed a gastrointestinal illness characterized by nausea, vomiting, diarrhea, and malaise.